数据多样性的理论研究[1]

陆彩女[1]　顾立平[2,3]　聂华[4]

([1]中国科学院上海药物研究所，上海，201203；[2]中国科学院文献情报中心，北京，100190；[3]中国科学院大学图书情报档案管理系，北京，101408；[4]北京大学图书馆，北京，100190)

摘要：数据多样性是数据的本质属性。在信息技术突飞猛进式发展和开放科学数据的时代背景下，数据多样性特征愈发明显。本文首先详细阐述数据多样性的内外表现，其中内部表现包括：科学数据生产过程的不同对象、数据出版的三位一体、不同学科采集暂存数据时不同的数据格式；外部表现包括数据生命周期加速了数据多样性、科研生命周期增加了数据多样性、数据在具体应用时被型塑而生发的多样性。随后，文章简要介绍了数据多样性的共同特征和影响因素，并从三个方面介绍了数据多样性的应用表征。对图书馆与馆员来说，认识数据多样性可以在一定程度上帮助科研人员解决数据汇交任务和数据披露压力，让数据重用变得简单并符合理想的数据生态体系。因此，作为一名数据馆员，需要有数据管理的能力并了解数据伦理的相关法律法规、政策与协议，努力为科研人员提供数据增值的业务。

关键词：数据多样性　科学数据　研究数据　数据服务

分类号：G350

Theoretical Research on Data Diversity

Abstract: Diversity is the essential attribute of data, especially scientific data. In the context of rapid development of information technologies (ITs) and the era of open research data, the characteristics of data diversity have become more obvious. Firstly, the paper elaborates the internal

---

[基金项目]本文系 2021 年度国家社会科学基金项目“开放科学环境中数据馆员服务模式研究”（项目编号：21BTQ005）的研究成果之一。（2021 National Social Science Fund Project “Data Librarian Service Models in open science environment”）

[通讯作者] 顾立平(ORCID:0000-0002-2284-3856)，博士，博士生导师，研究员，研究方向：科技信息政策、开放科学、开放获取、科研数据管理、用户服务等，Email：gulp@mail.las.ac.cn。

[作者简介] 陆彩女(ORCID: 0000-0003-0937-9312)，硕士，馆员，研究方向：开放获取、开放科学、科技信息政策、用户服务、计量分析等，Email：lucn@simm.ac.cn；顾立平(ORCID:0000-0002-2284-3856)，博士，博士生导师，研究员，研究方向：开放科学政策与法律咨询服务、数据科学理论与实践、用户工程，Email：gulp@mail.las.ac.cn；聂华(ORCID: 0000 0002 4522 5049)，研究馆员，研究方向：数字人文、人类信息行为与信息化研究等，Email: hnie@lib.pku.edu.cn。

and external manifestations of data diversity. The internal manifestations are different objects in the scientific data production process, the trinity of data publishing, and different data formats when colleting and depositing data in different disciplines. The external manifestations include data curation lifecycle accelerates the diversity of data, the research lifecycle increases the diversity of data, and diversity increased because of being sharpening when in practical use. Then, the paper describes the common features and impact factors of data diversity, and introduces the application representation of data diversity from three aspects. For libraries and data librarians, recognizing the diversity of data may probably help researchers solve the required task of data deposit and data release in open research data era, and making data reuse simple and creating an ideal data ecosystem. Therefore, as a data librarian, the data management capacity and the knowledge of relevant laws, regulations, policies, and agreements of data ethics are needed, in order to provide data value-added services.
Keywords: Diversity of data; Scientific data; Research data; Data services

1 引言

数据多样性，具有内在的三个特性和外在的三个特性。内在特性，是静态的，是数据作为一个对象或者物体与生俱来的特性。外在特性，是在它与环境和用户交互之后发生的，所以是动态的。数据多样性一直存在，但是未被发现、发觉和发展的原因，主要是在过去各个学科相对独立较少交叉，数据主要作为科研工作的附属物存在。但是到了数据驱动科研的时代，数据的价值和地位不断得到挖掘和重视，数据的多样性问题也越发凸显出来。如果忽视数据多样性，将会对新的数据的生产和管理带来阻碍。从数据交换迈入到数据开放的时代，从上而下，需要按照政策指导和要求对科学数据进行统一的管理和汇缴，甚至开放共享，进而出现了一系列的矛盾和问题。这些矛盾和问题，进一步促进了数据多样性的特点的凸显和发挥。数据安全以及数据交易的社会制度和相关政策的制定和变化，尤其在我国，也促进了数据多样性的发展。这最后一个原因尤其重要，因为美国和欧洲至今还未清楚认识到数据多样性，而我国图书馆界却对此早有认识，但是一直未能形成明确的概念并给出定义，直到笔者之一参与了书目多样性的研究之后，恍然大悟。

2020 年 4 月，中共中央、国务院发布《关于构建更加完善的要素市场化配置体制机制的意见》[2]，正式将“数据”作为一种新型的生产要素写入文件，并明确提出了加快培育数据要素市场发展的策略。数据要素市场化配置上升为国家战略，其重要性进一步凸显。2021 年 10 月，习近平在中共中央政治局第三十四次集体学习时强调，把握数字经济发展趋势和规律，推动我国数字经济健康发展[3]。数据资产化、数据产品、数据服务将是推动数据要素市

2 中国政府网.中共中央 国务院关于构建更加完善的要素市场化配置体制机制的意见[EB/OL](2020-04-09)[2021-09-30]. http://www.gov.cn/zhengce/2020-04/09/content_5500622.htm. (China government website. Opinions of the Central Committee of the Communist Party of China and the State Council on building a more complete system and mechanism for market-oriented allocation of factors[EB/OL](2020-04-09)[2021-09-30]. http://www.gov.cn/zhengce/2020-04/09/content_5500622.htm.)
3 新华社. 习近平：把握数字经济发展趋势和规律 推动我国数字经济健康发展[EB/OL](2021-10-19)[2021-10-20].https://mp.weixin.qq.com/s/46CT5gb-R9fnqn8ILHD72w. (Xinhua News Agency. Xi Jinping: Grasp the development trend and law of the digital economy to promote the healthy development of Chinese digital economy[EB/OL](2021-10-19)[2021-10-20].https://mp.weixin.qq.com/s/46CT5gb-R9fnqn8ILHD72w.)

场未来发展的重要力量。

2 科研数据开放共享与数据多样性理论的诞生

科研数据开放获取概念最早可追溯至 1950 年代，但是在最近十多年内才引起了人们的普遍关注和重视[4]。2003 年柏林宣言将科研数据作为学术知识的一部分并要求开放获取[5]，此后全球范围的国家政府机构、科研机构、科研资助机构、学术出版商等利益相关者都先后制定了科研数据开放共享政策[6,7,8,9,10,11,12]。随着全球越来越多的学术期刊要求开放共享论文底层数据，以及发表数据论文的数据期刊或混合期刊数量不断增加[13]，科研数据不再只是科研活动的副产品和附属物，而是已经逐渐成为科研活动的主要产品之一。同时，信息科技为科研数据共享提供了诸如数据存储、传输与处理等技术支撑，信息技术的不断发展也加快了科研数据共享的步伐。Elsevier 公司在其 2019 年发布的《科研的未来：下一个十年的驱动因素与场景》报告中指出，以信息技术发展为基础的科研数据开放共享将成为下一个十年科研活动最显著的特征，有望引发科研组织模式与科研创新模式的重大变革[14]。

科研数据开放共享的必然结果之一，是催生了数据多样性的概念和理论。从国家层面而言，数据资源已经是或即将成为一种新型生产要素，全球主要国家都在通过抢占科研数据开放共享的制高点，以尽可能争夺全球科研数据资源或保护本国科研数据资源不被收割。从科研机构、科研资助机构、学术出版商等利益相关者的角度来看，积极参与科研数据共享有助

---

4 盛小平,武彤.国内外科学数据开放共享研究综述[J].图书情报工作,2019,63(17):6-14. (Sheng Xiaoping, Wutong. Review on Open Sharing of Scientific Data Across the World[J]. Library and Information Science,2019,63(17):6-14.)

5 顾立平.数据级别计量——概念辨析与实践进展[J].中国图书馆学报,2015,41(02):56-71. (Ku Liping. Data Level Metric: Its Concepts and Progress[J]. Journal of Library Science in China,2015,41(02):56-71.)

6 张晓林.实施公共资助科研项目研究数据开放共享的政策建议[J].中国科学基金,2019,33(01):79-87. (Zhang Xiaolin. Policy Recommendations for Public Sharing of Research Data from Publicly Funded Research Projects[J]. Journal of Library Science in China,2019,33(01):79-87.)

7 阿儒涵,吴丛,李晓轩.科研数据开放的国际实践及对我国的启示[J].中国科学院院刊,2020,35(01):11-18. (A Ruhan, Wu Cong, Li Xiaoxuan. International Practice of Open Research Data and Its Enlightenment to China[J]. Bulletin of Chinese Academy of Sciences,2020,35(01):11-18.)

8 盛小平,武彤.国内外科学数据开放共享研究综述[J].图书情报工作,2019,63(17):6-14. (Sheng Xiaoping, Wutong. Review on Open Sharing of Scientific Data Across the World[J]. Library and Information Science,2019,63(17):6-14.)

9 尤霞光,盛小平.8 个国际组织科学数据开放共享政策的比较与特征分析[J].情报理论与实践,2017,40(12):40-45. (You Xiaguang, Sheng Xiaoping. Comparison of 8 International Organizations' Scientific Data Open Sharing Policies and Their Characteristics Analysis[J]. Information Studies: Theory & Application,2017,40(12):40-45.)

10 刘莉,刘文云,刘建,张宇.英国科研数据管理与共享政策研究[J].情报资料工作,2019,40(05):46-53. (Liu Li, Liu Wenyun, Liu Jian, etc. An Analysis of Scientific Research Data Management and Sharing Policy in the United Kingdom[J]. Information and Documentation Services,2019,40(05):46-53.)

11 赵瑞雪,赵华,郑建华,朱亮,寇远涛.科研机构科学数据管理实践与展望[J].农业大数据学报,2019,1(04):65-75. (Zhao Ruixue, Zhao Hua, Zheng Jianhua, etc. Scientific Data Management in Scientific Research Institutions: Practice and Prospects[J]. Journal of Agricultural Big Data,2019,1(04):65-75.)

12 马合,黄小平.欧美科学数据政策概览及启示[J].图书与情报,2021(04):84-91. (Ma He, Huang Xiaoping. Overview and Enlightenment of Scientific Data Policy in Europe and America[J]. Library & Information,2021(04):84-91.)

13 张晓林.实施公共资助科研项目研究数据开放共享的政策建议[J].中国科学基金,2019,33(01):79-87. (Zhang Xiaolin. Policy Recommendations for Public Sharing of Research Data from Publicly Funded Research Projects[J]. Journal of Library Science in China,2019,33(01):79-87.)

14 阿儒涵,吴丛,李晓轩.科研数据开放的国际实践及对我国的启示[J].中国科学院院刊,2020,35(01):11-18. (A Ruhan, Wu Cong, Li Xiaoxuan. International Practice of Open Research Data and Its Enlightenment to China[J]. Bulletin of Chinese Academy of Sciences,2020,35(01):11-18.)

于提升自己的影响力和话语权。对科研与学术生态体系而言，科研数据开放共享可实现科学研究的结果验证，可以丰富学术出版流程的标的物，并提升学术出版生态体系的良性循环。科研数据生产者、科研数据管理者、科研数据使用者在科研数据共享的趋势下，其主体多样性与数量增长趋势愈发明显。数据多样性原先只是具体科研工作或数据处理工作中需要面临的问题，但在开放科学和开放数据环境下，当我们需要制定相应的共享规范与原则或各种法律法规制度与政策时，或是基于科学数据提供相应的信息服务或数据服务与知识服务时，我们就需要用一个各方同意（认同）的框架来考虑和看待数据，此时数据多样性的问题就必然呈现和凸显出来。

3 数据多样性概念：内涵与特征

3.1 数据多样性内涵

3.1.1 数据多样性内在表现

数据多样性的内在表现包括以下几方面。

首先，数据本身就具有多样性的特征。在不同学科领域，数据集的形式不同，包括：社会科学经常使用具有变量和数值的试算表（sheet）数据、生命科学等经常描述组织结构的编码数据（code）、物理科学运用计算机进行模拟的模型数据（modeling）和以观测记录方式为主的科学学科的数字图像（image and voice record）等[15]。例如图 1 中的环境领域实测数据、微生物测序数据、蛋白质序列数据、高动态范围图数据相互之间对比来看都是各异的。

图 1 不同学科领域采集的数据[16]

其次，科研数据在数据集、数据描述、元数据的三位一体也是数据多样性的内在表现（如图 1），强调的是数据本身的生长与发展。其中，（1）数据集又称数据实体（Data Entity），是用来重复科研结果的证据；（2）数据描述是说明数据集的采集仪器、方法、产生过程、资助者等的描述性文档；（3）元数据就是描述数据集的贡献者（或生产者）、所属机构、所属学科、日期、版本等属性的信息[17]。元数据编目领域在用新 RDA 来描述图书或其他实体时

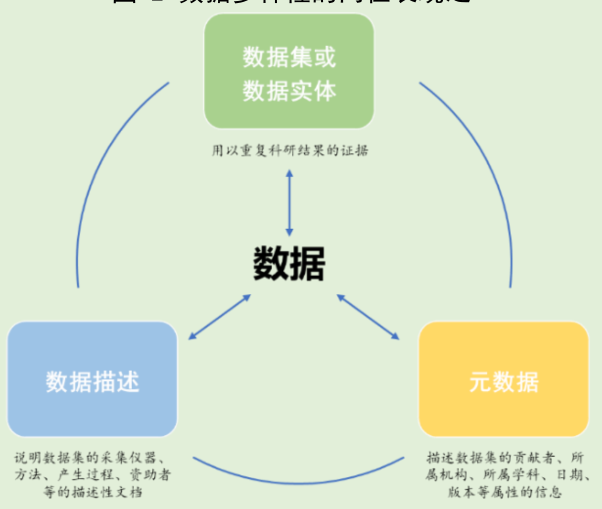
15 中国科学技术协会.中国科技期刊发展蓝皮书（2021）[M].北京：科学出版社, 2021. (China Association for Science and Technology. Blue Book of the Development of Chinese Sci-tech Journals(2021) [M].Beijing: Science Press,2021.)

16 本研究提供。

17 中国科学技术协会.中国科技期刊发展蓝皮书（2021）[M].北京：科学出版社, 2021. (China Association for Science and Technology. Blue Book of the Development of Chinese Sci-tech Journals(2021) [M].Beijing: Science Press,2021.)

也体现了数据多样性，即同一个 RDA，在编目时有不同的选项，如可以用本地控制词表代替工具包中提供的词表，从而使编目数据具有可扩展性[18]。

图 2 数据多样性的内在表现之一



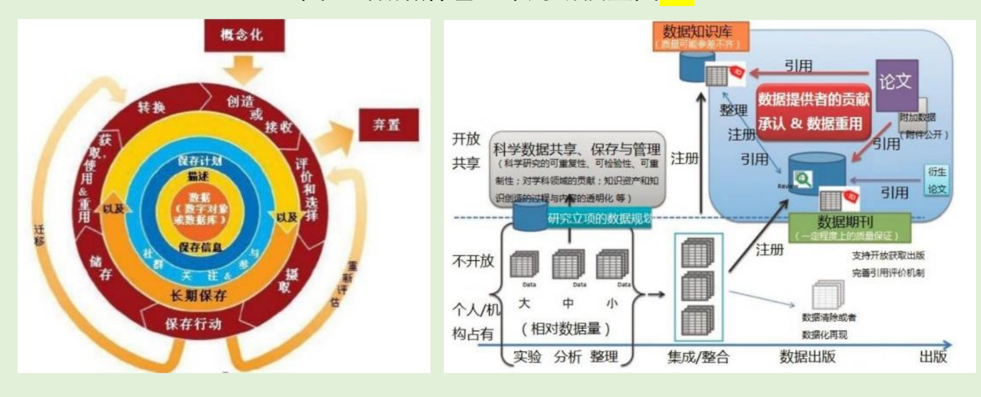
最后，从计算机处理数据的角度来看，数据本身也会因各种实际需要而变换成多种格式。例如同一组蛋白质组数据可以用序列、字典、元组、列表等方式来表示和存储；同一个物质也可以使用二维或三维结构图、分子式、结构简式、电子式等表示与存储。这也是数据多样性的一种内在表现形式。

3.1.2 数据多样性外在表现

数据多样性的外在表现主要强调数据的外在可扩展、可延伸、可应用的多样性。数据多样性的外部表现也具有以下三个方面。

首先，数据本身在其生命周期中或是在数据生产或重用时，会产生多个不同的版本，每个版本或多或少会并入新的数据（集）或是从中抽取子集，这也是一种多样性。不同的数据集、不同的数据版本，这就是数据多样性的第一种外部特征。可以说，数据生命周期加速了数据多样性。或者说，数据生命周期是数据多样性不断加速的过程。

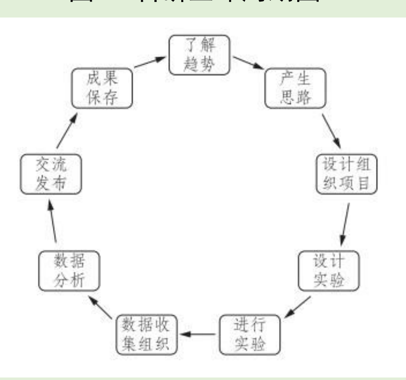
图 3 数据治理生命周期模型图[19][20]



18 胡小菁. 从数字化到数据化：编目的未来之路[EB/OL](2018-11-26)[2021-11-12].http://project.calis.edu.cn/jianbao/73/pdf/2.pdf. (Hu Xiaojing. From digitization to datalization: the future of cataloging[EB/OL](2018-11-26)[2021-11-12].http://project.calis.edu.cn/jianbao/73/pdf/2.pdf.)
19 Digital Curation Center. Curation Lifecycle Model[EB/OL][2021-11-19]https://www.dcc.ac.uk/guidance/curation-lifecycle-model.
20 林和弘.研究论文の影响度を测定する新しい动き!论文"位で即时かつ多面的な测定を可能とする

其次，科研生命周期增加了数据多样性。在了解趋势、产生思路、设计组织项目、设计实验、进行实验、数据收集组织、数据分析、交流发布、成果保存这一不断循环上升的科研生命周期中，不同的人、不同的生产者等会产生不同的数据，不同时间段的实验也会产生不同的数据。也就是说，数据的内外部环境会产生多样的数据。每个周期产生的数据也都不一样。例如，在不同加工阶段产生的不同数据，包括：仪器采集的原始数据（Raw Data）、经过抽取或者合并的衍生数据（Derived Data）、经过挑选具有验证结果的科研数据（Research Data）[21]。另外，倘若我们将广义科研生命周期中的所有涉及到的学术记录都视为一种数据，并且进行管理和保存的话，那数据本身也是不一样的。随着信息、数据成为了一种泛在形式，嵌入社会生产、生活、消费过程中，信息资源的内涵和边界也从文献信息扩展为数据、文献、实体等一切表现为数字化形态的存在[22]。此时，数据也就成了一种广义的、宽泛的含义，多样性也就随之而凸显。

图 4 科研生命周期图[23]



最后，在数据科学、大数据、人工智能、数据建模、智能数据等具体应用领域，为了适应不同机器和应用程序/软件的要求，数据势必会发生变化，要被型塑、塑造（sharp），那么这一过程会使得数据的多样性发生了另外一种改变，包括在存储格式上面的这种变化，以适应不同的机器或软件。

数据多样性内在表现和外在表现并不是孤立的两个方面，而是数据多样性的一体两面，两者相辅相成（如图 2）。首先，内在表现中元数据的属性值就是另一种形式的外在表现。数据多样性内在表现形式越清晰、规范，那么其被应用的可能性就越大，即数据多样性的外在表现也就越明显。其次，从所有科研数据整体来看，数据多样性的外在表现越丰富，那么数

Altmetrics[J].科学技术动向研究, 2013, 3/4:3, 20-28. (Hayashi Kazuhiro. A new movement to measure the degree of shadow of a research dissertation! Altmetrics that enables instant and multifaceted measurement at the dissertation "position"[J]. Research on Trends in Science and Technology, 2013, 3/4:3, 20-28.)

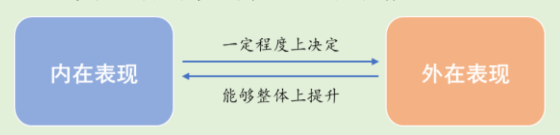
21 中国科学技术协会.中国科技期刊发展蓝皮书（2021）[M].北京：科学出版社, 2021. (China Association for Science and Technology. Blue Book of the Development of Chinese Sci-tech Journals(2021) [M].Beijing: Science Press,2021.)

22 刘细文.“创新开发科技信息资源，构建人工智能解决方案”报告[EB/OL](2021-10-15)[2021-11-19]. https://news.ruc.edu.cn/archives/348226. (Liu Xiwen. Innovatively develop scientific and technological information resources and construct artificial intelligence solutions[EB/OL](2021-10-15)[2021-11-19]. https://news.ruc.edu.cn/archives/348226.)

23 张晓林.研究图书馆 2020:嵌入式协作化知识实验室?[J].中国图书馆学报,2012,38(01):11-20. (Zhang Xiaolin. Research Libraries 2020: Knowledge Collaboratories? [J]. Journal of Library Science in China,2012,38(01):11-20.)

据生产的土壤和环境也会更加肥沃和健康，这就在一定程度上驱使更多的数据集及其数据描述与元数据的产生，数据多样性的内在表现也会更完整、统一且多样化。

图 5 数据多样性内外表现相互关系



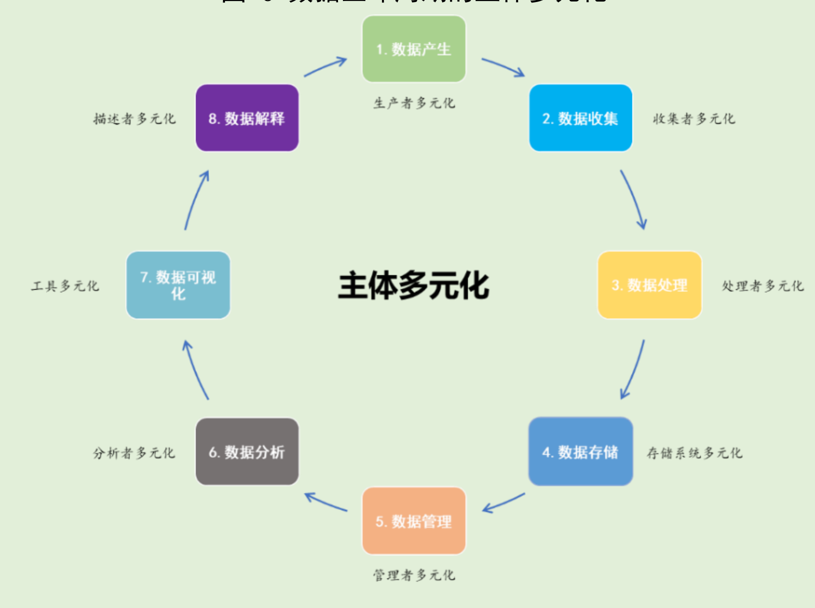
数据多样性是在科研过程和开放科学数据环境中形成的数据多种内外表现形式，是数据的本质属性之一，目的是为了实现数据的高效治理与应用而不断自我生长与自我发展。可以说，数据多样性既是手段，也是目的。

3.2 数据多样性共同特征

3.2.1 主体多元化

主体多元化是数据多样性的主要特征之一。从科研数据生命周期[24]来看，主体多元化几乎体现在数据生命周期的所有过程中。例如，在数据产生阶段，数据生产者是多元化的，而且不仅是生产者数量多、分布广，而且生产者类型多样，可能是观测机器、计算机，也可能是科研人员、实验员等。在数据存储阶段，数据可能存储在多样的系统中，包括可能存储在机构知识库、公共数据知识库、个人计算机等多个不同系统中。从内在表现来看，元数据、数据描述、数据集的维护主体同样也是多元化的。元数据的主体可能是图书馆和馆员，或是存储人员，数据描述和数据集的主体可能是科研人员或观测机器。

图 6 数据生命周期的主体多元化



3.2.2 协同发展

数据多样性另一个特征就是不同主体之间协同发展。科研人员产出数据集后，还需要数据管理（或治理）人员对数据的描述、元数据等做加工和处理，之后分析人员可能只选择一部分数据或子数据做分析和可视化，使用人员也会基于此科研数据产生新的数据或数据描

24 Stobierski T. 8 steps in the data life cycle[EB/OL](2021-02-02)[2021-10-19]. https://online.hbs.edu/blog/post/data-life-cycle?tempview=logoconvert.

述，以及正在形成或未来可能会形成的负责数据商品交易的交易员或交易平台以销售数据产品[25]等。可见，在数据生命周期中或科研数据生态体系中，扮演或发挥不同角色的主体之间相互协同、多元共治，共同推动科研数据多样性。

3.2.3 共同规则的制定

为了实现数据多样性，尤其是推动数据多样性内在表现的统一、完整，科学界、出版界、图书馆界等不同参与方之间应在相互协作的基础上，制定出共同的规则，推动数据的长效治理与高效利用。目前国际上已经出台了与科研数据相关的一系列规则和标准。例如，在元数据标准方面，全球已有约 65 个科研数据元数据标准[26]，其中常见的有：Dublin Core、数据文档计划 DDI、生态元数据语言 EML、地理空间领域的 ISO 19115 和 FGDC-CSDGM 等。数据管理与共享方面，有全球众多的组织和机构认可的 FAIR 数据共享原则[27]；在数据引用方面，数据出版和存储系统尽量为每条数据提供永久唯一标识符（persistent identifiers，简称 PID）或唯一标识符（Digital Object Identifier，简称 DOI）；数据引用原则和标准等[28][29][30][31][32]。目前，研究数据联盟 RDA 和世界数据系统 WDS 共同设立了学术链接交换工作组，努力制定论文-研究数据之间的关联规则并提供服务[33]。美国信息标准办公室 NISO 也宣布启动新项目来关联出版商与知识库之间的工作流，实现研究数据-论文之间的相互链接，针对元数据、术语、数据-论文关系的引用/链接类型等形成一系列标准或最佳实践[34]。

4 数据多样性影响因素

前文所述，数据多样性既是手段也是目的。那么，如何才能达到数据多样性这一目的呢？或者说怎么样去驱动并保持数据多样性？本文从以下三方面简要阐述。

4.1 环境因素

以自然环境为首的一系列环境，包括政治环境、经济环境、法律环境、科技环境等，是人类生存和生活的基础。从一定程度上而言，对于数据多样性，这些环境因素也是主要影响

25 Sands GE. How to build great data products[EB/OL](2018-10-30)[2021-10-20].https://hbr.org/2018/10/how-to-build-great-data-products.
26 Chen Sean, Alderete KA, Ball A. The RDA metadata standards directory[EB/OL][2021-10-23].http://rd-alliance.github.io/metadata-directory/standards/.
27 Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship[J]. Scientific Data, 2016, 3: 160018.
28 Data Citation Synthesis Group: Joint Declaration of Data Citation Principles[EB/OL](2014)[2021-10-23]. https://www.force11.org/datacitationprinciples.
29 ESIP Data Preservation and Stewardship Committee: Data Citation Guidelines for Earth Science Data, Version 2[EB/OL](2019-07-03)[2021-10-23].https://esip.figshare.com/articles/online_resource/Data_Citation_Guidelines_for_Earth_Science_Data_Version_2/8441816/1.
30 U.S. Geological Survey. Data Citations[EB/OL].[2021-10-24].https://www.usgs.gov/products/data-and-tools/data-management/data-citation.
31 Social Science Data Editors. Guidance on Data Citations[EB/OL][2021-10-25].https://social-science-data-editors.github.io/guidance/addtl-data-citation-guidance.html.
32 I4OC. Initiative for Open Citations[EB/OL][2021-10-25].https://i4oc.org/#about.
33 Research Data Alliance. RDA/WDS Scholarly Link Exchange (Scholix) WG[EB/OL][2021-11-16].https://www.rd-alliance.org/groups/rdawds-scholarly-link-exchange-scholix-wg.
34 The National Information Standards Organization. NISO Announces New Project to Integrate Publisher and Repository Workflows[EB/OL](2021-10-27)[2021-11-16]. http://www.niso.org/press-releases/2021/10/niso-announces-new-project-integrate-publisher-and-repository-workflows.

因素。尤其是开放共享的环境和各种完善的法律、科技、经济环境，都是数据多样化生产和使用的基础因素。此外，竞争环境，包括个人层面、机构层面、国家层面的竞争，也是数据多样性的重要因素。没有竞争，就可能会出现垄断局面，这势必不利于数据多样性。

4.2 技术手段

在信息化、数字化时代，数据通常以数字化的形式存储和展示，数字化的数据离不开数据库、网络与信息通讯技术（统称为信息技术）。在人工智能时代，人工智能技术能够影响数据多样化子集的推荐、大数据处理与分析、数据高速存储与传输等方面，进而也会影响数据多样性。此外，区块链、云计算等技术也会影响数据的存储与传输；数据分析与可视化技术会影响数据的应用与展示，这些技术都会从各个层次和角度影响数据多样性，尤其是数据多样性的外在表现。

4.3 标准遵从

如果不遵从标准规范，那么数据就不只是具有多样性，而是变得混乱无序，乃至无法被发现、被访问、被获取、被利用。科研数据相关的一系列标准，包括数据出版标准、数据引用标准、元数据标准、数据描述标准、数据使用标准及未来可能的研究数据-论文链接标准等，都是数据有序规则下多样化的保障手段。假设数据生态体系中的所有利益相关者（包括生产者、出版者、管理者、使用者、资助者等）都不遵从各种标准，那么数据生产者的权益无法保障，数据共享的方式无从知晓，甚至是未来可能出现的数据价值评估、数据产品/商品交易无法形成。

在《重新认识图书馆》[35]中指出：新型图书馆服务，应当具有：资源为基，技术为翼，需求为本，服务为王。本文认为对于数据多样性而言，环境为基，技术为翼，标准为本。三者结合，才能保证数据多样性的良性发展，以及在此基础上开展的数据应用和数据服务，甚至是基于数据、文献、知识等结合的情报服务、智库服务和智能服务等。当然，这里的标准为本不是说就“死守”已有各种标准，而是指各种原则、标准是保障数据多样性的原本或根本。

5 数据多样性理论的应用表征

数据多样性涉及多个学科领域，包括商业智能（Business Intelligence，简称 BI）、数据库、网络与信息通讯、数据出版、战略规划（数据策略）、数据模型（或数据建模）、数据治理、数据质量、数据素养、小数据应用、智能数据等。

5.1 数据战略中蕴藏的数据多样性

数据多样性体现在方方面面。在污染监测领域，数据多样性可以帮助企业遵守环境法规，数据科学家们通过从企业运营中捕获的环境数据，可以将其与其他运营数据一起分析，进而通过创建可操作的洞察力来提供企业竞争优势，提高业务效率[36]。不过真正多样化的数据驱

---

35 初景利. 重新认识图书馆[EB/OL](2021-10-18)[2021-10-20].http://t.cn/A6MWViP5. (Chu Jingli. Re-understanding the library[EB/OL](2021-10-18)[2021-10-20].http://t.cn/A6MWViP5.)
36 Data Diversity[EB/OL][2021-09-26].http://www.datadiversity.net/.

动战略，是要超越组织（或机构、企业）已有的现成数据或最容易收集的数据，从其主要活动和运营中立即可用的数据之外的数据中发现新东西[37]。例如，在营销领域，广告商通过分析他们的产品如何、何时和何地被谈论、拍照和发布到社交媒体，以更好地了解客户；在农业方面，农民已经习惯使用卫星和气象数据来确定作物的最佳时间和位置[38]。

5.2 数据多样性在大数据聚类中的应用

计算机科学家们也在积极探索、开发和利用数据多样性。例如，MIT 计算机科学和人工智能实验室联合 MIT 信息与决策系统实验室的研究人员就提出了一种基于多样性的新算法，保证从海量数据集中抽取样本子集时，各子集能保留完整集中的多样性特征[39]。这一算法可应用于各种推荐场景，如图书或电影等推荐；还可用于大规模学习中[40]。数据多样性这一属性在许多其他应用场景中也都发挥了关键的作用，例如基因网络子采样、文档提炼总结、视频摘要化、内容驱动搜索、推荐系统、传感器放置，及新闻标题或检索结果提示、影像或照片场景聚类、引文链研究方向识别、生物序列或多媒体数据聚类等[41]。

5.3 数据多样性在小数据领域的体现

数据多样性还体现在小数据领域。小数据尽管没有统一的定义，诊断数据、物种研究数据等都属于小数据。所以，某些科研数据也算是一种小数据。2021 年 9 月，美国网络安全和新兴技术局发布的研究报告《小数据人工智能的巨大潜力》中指出，小数据方法是一种只需少量数据集就能进行训练的人工智能方法，适用于数据量少或没有标记数据可用的情况，减少对人们收集大量现实数据集的依赖[42]。小数据方法包括迁移学习、数据标记、人工数据生成、贝叶斯方法、强化学习[43]，这些方法可用于图像识别、机器学习等领域。其中，迁移学习、数据标记和主动学习都符合前文所述数据多样性的特征。

6 开放科学数据环境下科研人员的痛点与图书馆及馆员的机遇

6.1 开放科学数据环境下科研人员的痛点

开放科学和开放数据环境下，科研人员的痛点和难点也越来越多。首先，科研人员和研

37 Marr B. Beyond big data: why data diversity is a crucial success factor[EB/OL](2017-12-19)[2021-10-26]https://www.qlik.com/blog/beyond-big-data-why-data-diversity-is-a-crucial-success-factor.
38 Marr B. Beyond big data: why data diversity is a crucial success factor[EB/OL](2017-12-19)[2021-10-26]https://www.qlik.com/blog/beyond-big-data-why-data-diversity-is-a-crucial-success-factor.
39 MIT's Computer Science and Artificial Intelligence Laboratory. Data diversity[EB/OL](2016-12)[2021-11-26]https://www.csail.mit.edu/news/data-diversity.
40 Li Chengtao, Jegelka S, Sra S. Fast Mixing Markov Chains for Strongly Rayleigh Measures, DPPs, and Constrained Sampling[EB/OL](2016-08-02)[2021-11-03].https://arxiv.org/abs/1608.01008.
41 Li Chengtao, Jegelka S, Sra S. Fast Mixing Markov Chains for Strongly Rayleigh Measures, DPPs, and Constrained Sampling[EB/OL](2016-08-02)[2021-11-03].https://arxiv.org/abs/1608.01008.
42 AI 科技评论.小数据,大前景!美国智库最新报告:长期被忽略的小数据人工智能潜力不可估量[EB/OL](2021-11-07)[2021-11-10].https://baijiahao.baidu.com/s?id=1715747682256222824&wfr=spider&for=pc.(AI Technology Review. Small data, big prospects! The latest report of the US think tank: the long-ignored small data artificial intelligence potential is immeasurable[EB/OL](2021-11-07)[2021-11-10].https://baijiahao.baidu.com/s?id=1715747682256222824&wfr=spider&for=pc.)
43 AI 科技评论.小数据,大前景!美国智库最新报告:长期被忽略的小数据人工智能潜力不可估量[EB/OL](2021-11-07)[2021-11-10].https://baijiahao.baidu.com/s?id=1715747682256222824&wfr=spider&for=pc.(AI Technology Review. Small data, big prospects! The latest report of the US think tank: the long-ignored small data artificial intelligence potential is immeasurable[EB/OL](2021-11-07)[2021-11-10].https://baijiahao.baidu.com/s?id=1715747682256222824&wfr=spider&for=pc.)

究团队需要应对越来越多的数据汇交任务，包括制定数据管理计划、开放数据、提交元数据、长期保存等，还要应对科研诚信、科研伦理和绩效考核等[44][45][46][47]。

图 7 科研人员需应对的数据任务与职责[48]



其次，数据披露已经成为科研人员的压力。在开放数据研究中发现，科研团队及其研究人员在面临数据披露时还要考虑一些外部资金、相关规范等，他们需要有人能为其提供全流程的数据咨询服务[49]，而不是简单的指南或最佳实践。

---

44 杨淑娟,陈家翠.研究成果传播与共享——英美国家基金项目数据管理计划概述[J].情报杂志,2012,31(12):176-179+69. (Yang Shujuan, Chen Jiacui. Dissemination and Sharing of Research Results: A Review of Anglo－American Funding Program Data Management Plan[J]. Journal of Intelligence,2012,31(12):176-179+69.)

45 王璞.英美两国制定数据管理计划的政策、内容与工具[J].图书与情报,2015(03):103-109. (Wang Pu. Promoting Long-term Preservation and Sharing of Research Data: Policies、Contentsand Tools[J]. Library & Information,2015(03):103-109.)

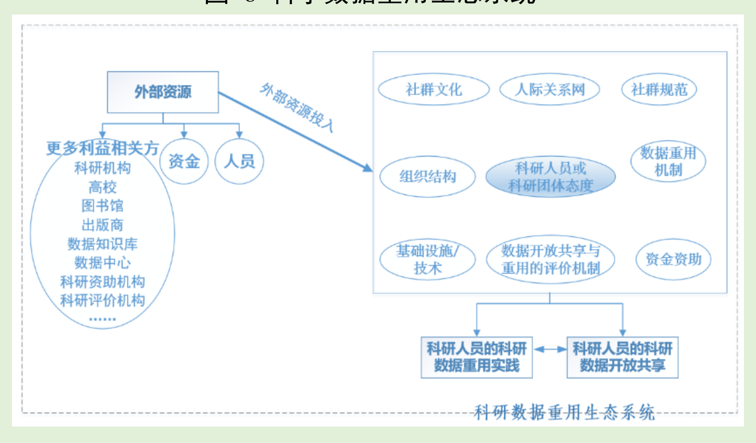
46 陈秀娟,胡卉,吴鸣.英美数据管理计划与高校图书馆服务[J].图书情报工作,2015,59(14):51-58. (Chen Xiujuan, Hu Hui, Wu Ming. Design and Application of Multi-dimensional Aggregation Based on Huizhou Culture Digital Resources[J].Library and Information Service,2015,59(14):51-58.)

47 彭鑫,邓仲华,李立睿.英国基金机构数据管理计划的实践调查与分析[J].图书情报工作,2016,60(17):27-32+39. (Peng Xin, Deng Zhonghua, Li Lirui. Practice and Enlightenment of LYRASIS － Typical Sample of Merger of Cross-system Library Consortia in U.S.[J].Library and Information Service,2016,60(17):27-32+39.)

48 山地一禎. RCOS はポジティブ思考な研究環境を提供できるか[EB/OL](2021-07-16)[2021-08-16].https://rcos.nii.ac.jp/diary/2021/07/20170716-1/index.php. (Sankanhi Ichiyo. Can RCOS provide a positive thinking research environment? [EB/OL](2021-07-16)[2021-08-16].https://rcos.nii.ac.jp/diary/2021/07/20170716-1/index.php.)

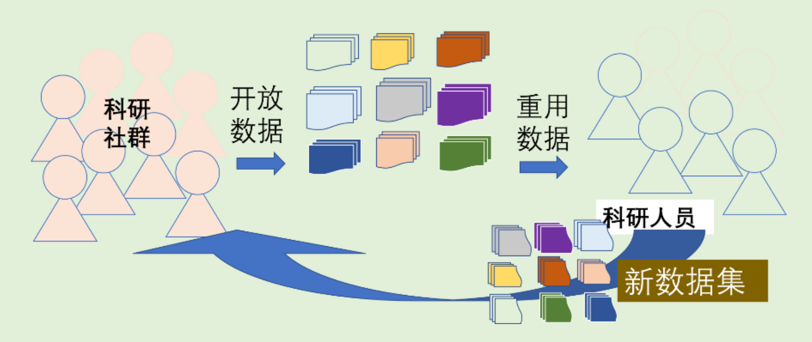
49 顾立平,张潇月.开放科学环境下数据馆员的实践探析[J].图书情报知识,2020(02):60-74+112. (Gu Liping, Zhang Xiaoyue. Exploration of Data Librarian Practice in Open Science[J].Documentation,Information & Knowledge,2020(02):60-74+112.)

图 8 科学数据重用生态系统[50]



最后，数据重用难以实现。数据重用的理想状态或者说理想的数据生态体系是，科研人员在利用开放数据后能产生新的数据或数据库并开放给其他人共享。但是经调研，有些科研人员在面对数据开放时就会变得犹豫和迟疑，这就使得数据重用变得困难[51]。

图 9 科研数据重用机制[52]



6.2 开放科学数据环境下数据馆员面临的挑战

馆员需要重新认识数据多样性，找到科研人员的数据痛点，以帮助科研人员解决上述问题和麻烦，不过也面临着一系列的挑战。

首先，数据馆员面临的挑战是数据管理的能力，包括存储、管理、汇交、保存的能力。当然，数据管理的能力需要信息基础设施的支持和数据馆员的业务支持。同时，在科研流程中，前、中、后等各个不同阶段所需要的数据支持也不一样。

50 张潇月,宋秀芳,顾立平,刘金亚,陈新兰.我国科研人员科研数据重用行为影响因素实证研究[J].情报学报, 2021. (Zhang Xiaoyue, Song Xiufang, Gu Liping, etc. An Empirical Study on the Influencing Factors of Scientific Research Data Reuse Behavior of Chinese Researchers[J]. Journal of the China Society for Scientific and Technical Information,2021.)
51 张潇月,宋秀芳,顾立平,刘金亚,陈新兰.我国科研人员科研数据重用行为影响因素实证研究[J].情报学报, 2021. (Zhang Xiaoyue, Song Xiufang, Gu Liping, etc. An Empirical Study on the Influencing Factors of Scientific Research Data Reuse Behavior of Chinese Researchers[J]. Journal of the China Society for Scientific and Technical Information,2021.)
52 制图：张潇月。

图 10 不同研究阶段馆员应具备的数据管理能力[53]

其次，馆员面临的第二个挑战是数据伦理的交流，包括法律、法规、政策和协议。作为数据馆员来说，应该了解著作权法[54]、数据安全法[55]、个人信息保护法[56]；以及数据管理办法[57]、出版管理条例[58]、电子出版物出版管理规定[59]；还应该了解相关的宏观政策，例如知识产权强国建设纲要[60]、学术期刊繁荣发展的意见[61]、人才强国战略[62]；以及知识共享（CC）许可

53 尾城孝一. 研究データ管理を担う人材育成のための教材開発[J].情報の科学と技術, 2019, 69(5):216-218. (Koichi Oshiro. Development of teaching materials for human resource development responsible for research data management[J].Information science and technology,2019, 69(5):216-218.)
54 中华人民共和国著作权法（2020 年国家主席令第 62 号）[EB/OL](2020-11-11)[2021-11-23]. http://scjgj.yibin.gov.cn/sy/xxgk/zcwj/202011/t20201112_1379367.html. (Copyright Law of the People's Republic of China (2020 Presidential Order No. 62) [EB/OL](2020-11-11)[2021-11-23]. http://scjgj.yibin.gov.cn/sy/xxgk/zcwj/202011/t20201112_1379367.html.)
55 中华人民共和国数据安全法[EB/OL](2021-07-30)[2021-11-23].http://ggzy.gzlps.gov.cn/zfxxgkml/zcfg/202107/t20210730_69354088.html. (Data Security Law of the People's Republic of China[EB/OL](2021-07-30)[2021-11-23].http://ggzy.gzlps.gov.cn/zfxxgkml/zcfg/202107/t20210730_69354088.html.)
56 中华人民共和国个人信息保护法[EB/OL](2021-08-20)[2021-11-23]. http://www.gov.cn/xinwen/2021-08/20/content_5632486.htm. (Personal Information Protection Law of the People's Republic of China[EB/OL](2021-08-20)[2021-11-23]. http://www.gov.cn/xinwen/2021-08/20/content_5632486.htm.)
57 国办. 国务院办公厅关于印发科学数据管理办法的通知[EB/OL](2018-04-02)[2021-11-23]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm. (General Office of the State Council. Notice of the General Office of the State Council on Issuing the Measures for the Administration of Scientific Data[EB/OL](2018-04-02)[2021-11-23]. http://www.gov.cn/zhengce/content/2018-04/02/content_5279272.htm.)
58 国家新闻出版署.出版管理条例(2016 修正队)[EB/OL](2017-12-26)[2021-11-23].http://www.nppa.gov.cn/nppa/contents/309/5966.shtml. (National Press and Publication Administration. Publication Management Regulations (2016 Revision)[EB/OL](2017-12-26)[2021-11-23].http://www.nppa.gov.cn/nppa/contents/309/5966.shtml.)
59 电子出版物出版管理规定[EB/OL](2008-02-21)[2021-11-23].http://www.gov.cn/gongbao/content/2009/content_1388688.htm. (Regulations on the Administration of Electronic Publications[EB/OL](2008-02-21)[2021-11-23].http://www.gov.cn/gongbao/content/2009/content_1388688.htm.)
60 知识产权强国建设纲要（2021－2035 年）[EB/OL](2021-09-23)[2021-11-23]. http://www.cnipa.gov.cn/art/2021/9/23/art_2742_170305.html. (Outline of Building a Powerful Country with Intellectual Property[EB/OL](2021-09-23)[2021-11-23]. http://www.cnipa.gov.cn/art/2021/9/23/art_2742_170305.html.)
61 国家新闻出版署.关于推动学术期刊繁荣发展的意见[EB/OL](2021-05-18)[2021-11-23]. http://www.nppa.gov.cn/nppa/contents/312/76209.shtml. (National Press and Publication Administration. Opinions on Promoting the Prosperity and Development of Academic Journals[EB/OL](2021-05-18)[2021-11-23]. http://www.nppa.gov.cn/nppa/contents/312/76209.shtml.)
62 中国共产党新闻网. 人才强国战略是实现国家强盛的第一战略[EB/OL](2017-11-02)[2021-11-23].http://theory.people.com.cn/n1/2017/1102/c40531-29623743.html. (Cpcnews.cn. The strategy of strengthening the country by talents is the first strategy to realize the prosperity of the country[EB/OL](2017-11-02)[2021-11-23].http://theory.people.com.cn/n1/2017/1102/c40531-29623743.html.)

协议[63]、自由软件许可[64]、数据库使用协议等。

最后，数据馆员在数据作为生产要素的要求下，可能还需要了解数据增值的业务，包括交办、交换、交易、交涉。尽管目前已经得到广泛应用的大都是通信、电商领域的数据[65]，但相信在不久的未来，科学数据领域也会展开数据交易，科研数据的数据确权问题也将浮现。

图 11 分类分级数据产权内容[66]

	个人数据	企业数据	社会数据
公共品	公有产权	公有产权	公有产权
准公共品	公有产权 基础数据产权	公有产权 衍生数据产权	公有产权 衍生数据产权
私有品	基础数据产权	公有产权 衍生数据产权	

6.3 与图书馆发展相关的资源系统建设与服务

图书馆作为信息资源的收藏、传阅、服务场所，在万物皆是数据的时代，跳出传统知识资源的界限已成为必然[67]。图书馆在描述资源、提供访问和建立馆藏以及为数字资源的长期管理提供支持方面有着悠久的传统[68]，部分图书馆也已经开始参与数据开发、整合和利用的全生命周期，并在更广泛的使命和服务范畴内呈现和分析[69]。从数据治理的角度而言，社会直接面向数据，数据直接影响社会，而图书馆居于其中的角色，更多不是中介，而是驱动者、促进者，以及辅助者的角色，图书馆可以充分应用已有的文献领域的经验积累，从技术、法律、伦理等规则进行引导。从知识服务的图书馆学理论而言，结合实践经验论证理论以及需要理论指导实践等的角度，都需要数据多样性，作为数据服务的一个理论支撑，因为在文献服务、信息服务、情报服务之后，数据服务是知识服务的最后一块拼图。在数据-信息-情报-决策-评价的情报价值链中，数据应被作为情报工作的起点[70]。

当前的开放科学生态体系，已从第一代的文献知识库和数据知识库为用户存储、检索和

63 Creative Commons[EB][2021-11-23].https://creativecommons.org/.
64 Free Software Foundation. Free Software Licensing Resources[EB/OL](2006-11-06)[2021-11-23]. https://www.fsf.org/licensing/education.
65 中国信息通信研究院.数据价值化与数据要素市场发展报告(2021 年). (China Academy of Information and Communications Technology. Data Valuation and Data Elements Market Development Report(2021).)
66 中国信息通信研究院.数据价值化与数据要素市场发展报告(2021 年).(China Academy of Information and Communications Technology. Data Valuation and Data Elements Market Development Report(2021).)
67 储节旺,李振延.图书馆大数据知识生态系统特征及构成研究[J/OL].情报理论与实践:1-12[2021-10-27]. http://kns.cnki.net/kcms/detail/11.1762.g3.20210909.1452.002.html. (Chu Jiewang, Li Zhenyan. Research on the Characteristics and Constitution of Library Big Data Knowledge Ecosystem[J/OL].Information Studies: Theory & Application[2021-10-27]. http://kns.cnki.net/kcms/detail/11.1762.g3.20210909.1452.002.html.)
68 LIBER. Implementing FAIR data principles: the role of libraries[EB/OL](2017-12-08)[2021-11-10].https://libereurope.eu/article/implementing-fair-data-principles-role-libraries/.
69 蔡迎春,欧阳剑,严丹.基于数据中台理念的图书馆数据服务模式研究[J/OL].图书馆杂志:1-11[2021-10-27].http://kns.cnki.net/kcms/detail/31.1108.g2.20210825.1808.010.html. (Cai Yingchun, Ouyang Jian, Yan Dan. Research on Library Data Service Mode Based on the Concept of Data Center[J/OL].Library Journal[2021-10-27]. http://kns.cnki.net/kcms/detail/31.1108.g2.20210825.1808.010.html.)
70 刘细文.情报学范式变革与数据驱动型情报工作发展趋势[J].图书情报工作,2021,65(01):4-11. (Liu Xiwen. Paradigm transformation of library and information science and trends of data-driven information services[J].Library and Information Service,2021,65(01):4-11.)

使用，走到了第二代文献和数据之间的引用关联、元数据关联和第三方词表关联的数据产品阶段[71]。通过标准规范的互操作性，而非元数据的互操作性，第三代的开放科学生态体系，正在构建软件、代码、数据、文献、引用、评价内容等的“有机生长体”[72]。目前，数据领域以及文献领域所形成的超大元数据集成，正在朝向类似的数据产品的方向发展。初代的数据产品原型有：数据、数据集、元数据、关联数据、语义数据、开放政府数据、研究数据、数据论文与数据出版等。

图书馆作为资源集成体，在数据资源建设规划中，就应考虑数据多样性，使得数据能够尽可能地为后来的人所使用，尽可能地提升数据的使用价值。对于已有的数据资源，也要考虑数据多样性，即数据如何能够尽可能地在各种情景下被使用。如果在数据资源建设规划阶段没有考虑数据多样性，或者边建设边规划，那么由于数据不同于文献，一旦建好了之后，就会有其限制或无法被更好地使用。反之，如果在数据多样性原则的指导下建立数据资源，且能被用户使用并有利于科研，那么就能不断地开展数据资源开发与利用的良性循环。数据多样性旨在提升数据的可发现，可获得，可交互及可重用，其与研究数据的 FAIR 原则在内涵有相通之处。图书馆重视数据多样性也能在一定程度上确保研究数据遵循 FAIR；反之，图书馆从遵循研究 FAIR 原则出发，也能保证使用通用的元数据体系或编码体系来描述、注释、归档研究数据，也就增强了数据的多样性。

作为馆员或数据馆员来说，从事馆藏知识数据库建设、管理与推广政策已经成为一项主要职责[73]，在此过程中，探索和制定元数据标准即最佳管理实践，注重数据质量、可获取性、互操作至关重要。馆员可以尝试在数据管理计划实践中，认识到数据多样性的重要性，并通过提升数据采集、描述、整理、存储过程中的方法、政策、标准等的完备性，保障研究数据的多样性。此外，馆员还可以提供嵌入式数据支持服务，帮助科研人员制订数据计划、整理和处理数据、分析数据并可视化、保存数据等，为数据使用者和生产者提供无缝对接的配套服务，这也从外在表现提升了数据多样性。馆员在信息组织领域拥有丰富的经验，还能积极利用这些信息组织经验，转移到数据领域（尤其是转移到科研数据或小数据管理领域），积极参与并努力做好数据描述、数据标记或数据编目，为智能情报系统提供更好的数据加工和管理服务。

7 结语

多样性意义重大、影响深远。一切形式的文化多样性都是与经济繁荣息息相关的竞争差

71 中国科学技术协会.中国科技期刊发展蓝皮书（2021）[M].北京：科学出版社, 2021. (China Association for Science and Technology. Blue Book of the Development of Chinese Sci-tech Journals(2021) [M].Beijing: Science Press,2021.)

72 中国科学技术协会.中国科技期刊发展蓝皮书（2021）[M].北京：科学出版社, 2021. (China Association for Science and Technology. Blue Book of the Development of Chinese Sci-tech Journals(2021) [M].Beijing: Science Press,2021.)

73 刘红菊,陈阳.美国高校图书馆数据馆员岗位研究及思考[J].图书馆建设,2019(01):135-140+146. (Liu Hongju, Chen Yang. Research and Thinking on the Posts of Data Librarians in American University Libraries[J].Library Development,2019(01):135-140+146.)

异化因素[74]。数据多样性，作为一种文化多样性，在数据时代只有被真正认识和努力实现，这样的组织才能更好地适应新思想、新技术以及新的社会和经济挑战。从图书馆和馆员角度来说，数据多样性是图书馆和馆员提供数据服务的基石，也是数据情报工作的起点，同时也是图书馆和馆员深入参与数据驱动科学发现的发展机遇所在。

74 Adams J, Pendlebury D, Szomszor M. 研究布局中的学科多样性：概念、测度及其在创新活动中的作用[EB/OL](2021-08)[2021-11-10].https://img02.ma.scrmtech.com/18476/1812/resource/1629167704/%E7%A0%94%E7%A9%B6%E5%B8%83%E5%B1%80%E4%B8%AD%E7%9A%84%E5%AD%A6%E7%A7%91%E5%A4%9A%E6%A0%B7%E6%80%A7.pdf. (Adams J, Pendlebury D, Szomszor M. Disciplinary diversity in research layout: concept, measurement and its role in innovation activities[EB/OL](2021-08)[2021-11-10].https://img02.ma.scrmtech.com/18476/1812/resource/1629167704/%E7%A0%94%E7%A9%B6%E5%B8%83%E5%B1%80%E4%B8%AD%E7%9A%84%E5%AD%A6%E7%A7%91%E5%A4%9A%E6%A0%B7%E6%80%A7.pdf.)